Arabic Text Classification: The Effect of the AWN Relations Weighting Scheme

Suhad A. Yousif, Venus W. Samawi, *Member, IAENG,* and Islam Elkabani

***Abstract*— Text Classification (TC) is a major challenge in terms of the Arabic language given its diacritics, which affect the meaning of words. The considerable growth of Arabic documents found on the web necessitates proper Arabic TC techniques. Arabic Text Classification (ATC) requires extensive work to analyze the content of Arabic documents, where matching by bag of words (BoWs) alone may be insufficient. In this research, the BoWs and semantic relations between words are used to enhance the accuracy of ATC. These relations are constructed using Arabic WordNet (AWN) thesaurus as a lexical and semantic source. According to the depicted results, some of these relations showed more effectiveness than others that barely improved the classification accuracy. Therefore, a weighting scheme is suggested to assign different weights to each relation based on the frequency of the relation in the AWN and the corpus. This approach aims to construct a training file, which includes BoWs and the corresponding relation terms extracted from AWN with their weights. The Naive Bayes algorithm is used as a classifier. Moreover, the performance of the suggested approach is measured using the F1-measure. The variability of results is reduced through the k-fold cross-validation technique, with K=10. Experimental results show that ATC using the weighting scheme approach outperforms the BoWs approach.**

***Index Terms*—Text Classification, Natural Language Processing, AWN, Semantic Relations, Naive Bayes**

I. Introduction

TEXT classification (TC) is the process of assigning a text document (or a set of documents) to one or more predefined classes based on the document content [1, 2]. This process is an important field of study [3]. The massive amount of text documents, available on the World Wide Web (WWW), complicated the process of information retrieval especially for text documents lacking keywords (particularly old documents). Furthermore, although some text documents contain keywords, such keywords do not express the full content of the document, which hinders document retrieval. Automatic text classification systems (ATCS) are used to help in locating and retrieving such documents correctly and rapidly. Thus, improving the accuracy of automatic text classifiers is an essential issue in enhancing text classification process.

ATCS are utilized in various applications, such as email-filtering, Web page and automatic article indexing, document clustering, and natural language processing. The expansion of the TC method for Arabic documents is a challenging problem because of the complexity and the nature of the Arabic language.

Arabic language is one of the supreme languages, which is used by nearly 300 million people. Therefore, an increasing concern arises for developing efficient methods for processing Arabic language. In their attempts, most researchers for Arabic language utilized a statistical method, namely, the bag of words (BoWs) method [4]. With BoWs, the document is classified into a category based on the ratio of mutual words among documents that belong to the same category. The selected feature is represented by the frequency of each word in the text document [3, 5–10]. This feature expresses deficiency semantic information to classify text accurately.

A minimal number of studies use the lexical, semantic, and lexio-semantic relations of the Arabic WordNet (AWN) thesaurus or other thesauruses, such as Wikipedia. They rely on using the synonym of words (concept) [7, 11] or on using the different semantic relations between words that are available inside the AWN, such as hyponym, related to, verb group, and others. Recently [1, 9, 12] investigated all the relations between words available in the AWN thesaurus to enhance the accuracy of TC.

AWN contains words with their classes (nouns, verbs, adjectives, or adverbs), roots, and concepts (synsets), in addition to the relations among these concepts. These relations provide semantic information among concepts and their original words. In [9, 12], these concepts and their relations are exploited to improve Arabic TC. Therefore, some relations revealed more effectiveness than others (which could affect the classification capability in different ways). Accordingly, we suggest assigning different weights to the semantic relations based on their frequency in the corpus and the AWN. The performance of the proposed approach is evaluated using the F1-measure and compared with the BoWs approach. This approach aims to construct a training file, which includes BoWs and the corresponding relation terms extracted from the AWN with their weights.

Unlike English, no benchmark dataset exists for the Arabic language. Most researchers depend on collecting datasets from free online magazine news available on the Internet. Documents in this work are gathered from the BBC dataset. The BBC dataset is the most extensively used of its

Manuscript received March 21, 2017; Revised April 10, 2017; Accepterd on April 20,2017.

Suhad A. Yousif is with Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq (corresponding author Email: suhad.alezzi@gmail.com; say@sc.nahrainuniv.edu.iq).

Venus W. Samawi is with Department of Computer Multimedia Systems, Faculty of Information Technology, Isra University, Amman, Jordan (Email: venus.samawi@iu.edu.jo; dr_venus2004@yahoo.com)

Islam Elkabani is with Mathematics and Computer Science Department, Faculty of Science, Beirut Arab University, Beirut, Lebanon, (E-mail islam.kabani@bau.edu.lb), on leave from Mathematics and Computer Science Department, Faculty of Science, Alexandria University, Alexandria, Egypt.

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

kind. It is freely accessible, public, and proves sufficient for the classification process. The BBC dataset consists of seven predefined classes: Middle East news (MEN), world news (WN), collection, economies, magazine, sport, and technology, with 4763 documents and 1,860,785 words in all categories [5, 13]. Naive Bayes algorithm is used as a classifier. Finally, the performance of the suggested approach is measured using the F1-measure.

The rest of this paper is organized as follows. Section two explains the suggested ATC, the main model. Section three focuses on the proposed weighting method. The experimental results are assessed in section four. Finally, conclusions and future work are expressed in section five.

II. Suggested ATC: Main Model

In this work, a supervised text classifier is used (Naive Bayes classifier). The proposed classifier consists of two phases (the most important phases in any supervised text classifier):

- ***Training Phase***, which is used to produce the knowledge source (training file)? The training file contains BoWs and the corresponding relation terms extracted from the AWN with their weighted frequency.
- ***Testing Phase***, wherein the training file will be used to classify the document.

Fig. 1 illustrates the main TC phases and stages.

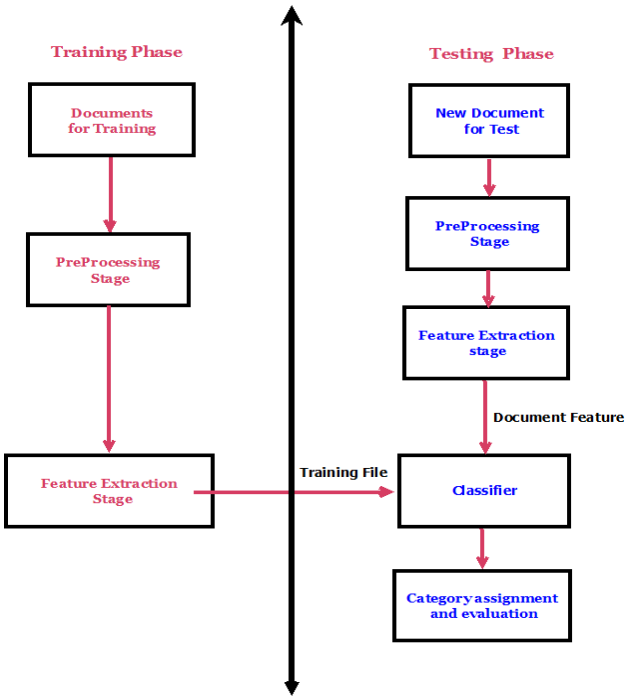

Fig. 1 Main Classification Phases

In Fig. 1, the training phase in the suggested ATC consists of three stages: preprocessing, feature extraction and selection, and training file construction. For the testing phase, the three stages include: preprocessing, feature extraction and selection, and the classification stage based on the training file. The classifier stages are:

A. Preprocessing Phase

In this phase, all irrelevant information affecting the accuracy of TC is removed. Such irrelevant information could be numbers, non-Arabic words, punctuation marks, redundant word, and stop words (prepositions and pronouns). This phase reduces both the dimensionality in the generated training file and the words from the document requiring classification. Next, the normalization process is implemented by swapping letters (" أ إ آ") with ("ا"), the letter ("ؤ ء ") with ("ا"), and finally, the letter ("ى") with ("ا"). Finally, only important words (features) are saved in the generated training file [12].

B. Feature Extraction

Mainly, there are two forms of feature extraction are represented in this work. The first form is the external feature that depends on the title, author name, publication date, author gender, and any other feature distinct from the content. The second feature is the internal feature that manages text document content and presents its linguistic features, such as lexical information and grammatical group [14, 15]. In this work, we deal with the internal features in two ways: by using a BoWs form as a another set of features and by using resulting concepts and semantic relations through utilizing the AWN lexical thesaurus as a set of features.

BoWs Lexical Feature Representation

BoWs are the simplest feature representation of text, in which all document texts are represented as a vector of words. Two significant scarcities, namely, polysemy and synonymy, exist in this representation. These scarcities occur because of the ambiguity of words and insufficient information about the relations of words. To redress these deficiencies, we suggest using a conceptual representation and semantic relations extracted from the AWN as a feature set.

Semantic Relations in AWN

ATC was estimated using various statistical methods that sometimes suffer from relatively low classification accuracy rate. We believe that classification accuracy could be improved using the semantic relationship between document terms and their related concepts using WordNet (a major enabler in the field of semantics), which can be viewed as a lexical thesaurus. WordNet has several versions belonging to different languages, including Arabic [8, 16].

AWN is a lexical database of the Arabic language. It comprises nouns, verbs, adjectives, and adverbs that are grouped into sets of cognitive synonyms (synset). Synsets are connected using semantic and lexical relations, which renders the structure of AWN as a useful tool for linguistics, TC, and natural processing [10, 17]. AWN groups words based on their meaning and connects them according to specific scenes. It has four tuples (tags):

- Item: The concepts of terms
- Word: The terms (words)
- Form: The root of the words
- Link: The relationships between concepts

The connections between the four tuples are required to extract information from AWN. Fig. 2 illustrates the three connections among the four components [18].

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

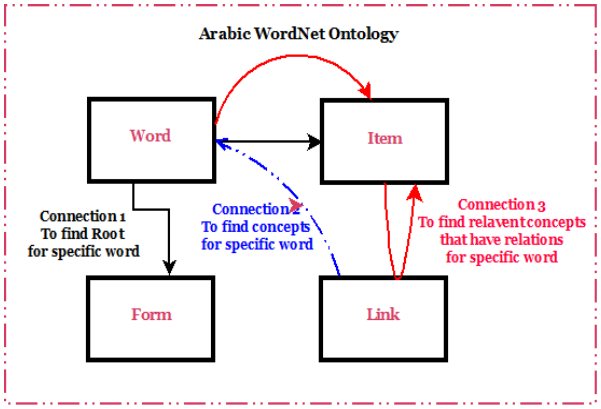

Fig. 2 Connections among the components of AWN thesaurus

The three connections illustrated in Fig. 2 are as follows.

- ***Connection 1*** (from word to form): the root of the individual word (term) is obtained by this connection.
- ***Connection 2*** (from word to item): the particular concept(s) synsets obtained by this connection are used to obtain the word. Each word may have more than one synset connected to it. For example, Aleph الف involves three concepts (شكل Shakkala كون Kawanaكتب ، Kataba) [6, 7].
- ***Connection 3*** (from word item to link item): the lists of related synsets that are relevant to the specified word are established by this connection. Fig. 3 shows the list of all relations found in the AWN ontology. For example, applying related_to relation *from the BBC dataset, we obtain:*

فرق، بضع، حز، اسقط، نشر،نقش، حفر، نمنم، شظى، ثلم، ثقب، نقب، خدد، تشويه، مسخ، جراحة، عملية، عملية جراحية، قص

Fasala, Faraqa, Shaqa, Batha'a, Asqata, Nashara, Naqasha, Hafara, Namnama, Shatha, Thalama, Thaqaba, Naqaba, Khadada, Tashweeh.

Fig. 3 represents the relations existing in AWN, which constitute the foundation of the proposed method.

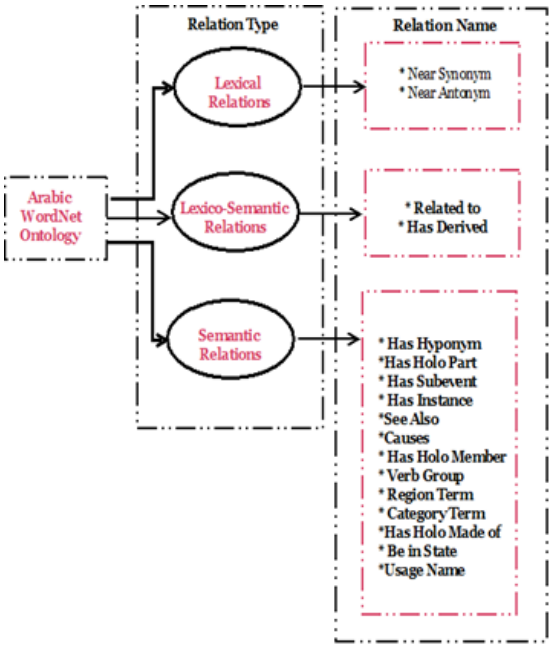

Fig. 3 Relation and Their Types in the AWN Thesaurus

III. Relations: The Weighting Method

The proposed weighting method aims to assign a different weight to each relation depending on the AWN thesaurus. Different weights are assigned to relations based on their frequencies in the AWN as displayed in Table I. Some relations, such has_hyponym, have high frequencies, which, according to the suggested approach, are given higher weight.

TABLE I
FREQUENCIES OF EACH RELATION IN THE AWN THESAURUS

Relation_Name available in AWN	Frequency of Occurrence in AWN	Relation_Weight
verb_group	152	0.0082
has_holo_member	334	0.0180
see_also	192	0.0104
usage_term	3	0.0002
has_hyponym	9352	0.5049
has_subevent	128	0.0069
be_in_state	83	0.0045
has_holo_madeof	60	0.0032
related_to	4774	0.2577
near_synonym	122	0.0066
has_derived	178	0.0066
has_holo_part	697	0.0376
has_instance	1067	0.0576
near_antonym	722	0.0390
Causes	75	0.0040
region_term	35	0.0019
category_term	548	0.0296
Total	18522	1.00

Given the inadequacy of AWN with regard to covering all words in the Arabic language, the relations of AWN cannot be entirely relied upon in weight calculation. Hence, we intersect between the original AWN thesaurus database and the BBC dataset to obtain the frequencies of the crossed words and its relation and use them in calculating the new weight.

TABLE II
EXAMPLES EXTRACTED FROM THE BBC DATASET.

BBC-Dataset	Frequency of Occurrence	% of Relation Occurrence
Verb-group	932	3.20%
Has_holo_part	551	2%
See_also	1129	4%
Usage_term	1	0%
Has_hyponym	10175	36.02%
Be_in_state	560	2%
Has_subevent	840	3%
Related_to	6171	21.80%
Has_holo_made_of	114	0.5%
Near_synonym	450	1.6%
Has_Derived	708	2.5%
Has_holo_part	1717	6%
Has_Instance	862	3%
Near_antonym	2078	7.3%
Causes	596	2.10%
Region_term	51	0.10%
Category_Term	1313	4.60%
Total Relations	28248	

The proposed method suggests adding a list of related words extracted from a specific original term by obtaining their relations to the new training file. Not all relevant words are added to the file. The related words are added depending on their final weight as calculated in Table III. According to

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the percentage presented in Table II, which is synchronized with the Table I values, we assign a high weight to the high-frequency occurrence of relation in the BBC dataset (i.e., has_hyponym and related_to relations have the highest frequency **F** fetched from Table II). Eq. (1) explains the computation of the weight of each document term in Table III.

$$W_{ij} = Freq(word_i, doc_j) \times AWN\_Weight(Relation(word_i)) \quad (1)$$

Example: the weight of the relation of has_hyponym to the $word_i$ = (عرضArath) in a doc_j from the BBC dataset is found using Eq. (1):

Relation weight (has_hyponym) = 0.504 from Table I.

Freq. = 202 (frequency of occurrence of $word_i$ in doc_j for has_hyponym relation in BBC) from Table II.

$W(word_i,doc_j)$ = 0.5049 × 202 = 101.99, which represents the new final weight as in Table III.

TABLE III
COMPUTING FINAL WEIGHT OF REAL EXAMPLE FROM THE BBC DATASET

Original word (term) =".فيلم" In MEN					
Related Word	**Relation Name**	**Class Name**	**AWN–Weight (Relation)**	**Freq.**	**Final weight**
عرض(Arath)	has_hyponym	MEN	0.5049	202	101.99
صور(Swaur)	related_to	WN	0.2577	117	30.15
مشهد (Mashhad)	has_holo_part	MEN	0.0376	57	2.14
استعراض (Istirath)	has_hyponym	WN	0.5049	8	4.03
فن (Fin)	category_term	MEN	0.0296	6	0.177
معرض (Ma'rah)	related_to	---	0.2577	0	0

This method can be explained as follows. A weight to each relation is assigned depending on the AWN thesaurus. Then, final weights can be computed by multiplying the relation weight by the frequency of the related word as depicted in Table IV. Subsequently, we attempted to select the only operative relation to reducing the size of the training file. Hence, a specific threshold value is specified by calculating the median of the final weight extracted from Table III. Eq. (2) expresses this process, in which the first part calculates the median when an odd number of words exists in the training file, while the second part calculates the median for an even number of words in training file. We construct the new training file consisting of only the related words with the threshold greater than the median value.

$$Median = \left(\frac{n+1}{2}\right)^{th} term$$

$$Median = \frac{\left(\frac{n}{2}\right)^{th} term + \left(\frac{n}{2}+1\right)^{th} term}{2} term \quad (2)$$

In the example of Table III, the threshold value (median) is 3.08. Only the three related words (عرض aratha, صور Swaur, استعراض istirath) will be selected because their final weight is > 3.08. Related words with a final weight less than the threshold will be ignored (فن fin, مشهد mashhad, ma'rath معرض). The selected related words will be added to the original term in the new training file.

For the whole BBC dataset, the calculated threshold value using median equals to 2, and the training file will contain all original words with any relation that has a final weight with a value >2.

IV. EXPERIMENTAL RESULTS: ASSESSMENT

The simplest and most effective learning algorithm used for ATC is the Naive Bayes algorithm (NB) [19–21], which is demonstrated in Eq. (3).

$$C_{NB} = \arg\max P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

$$P(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|} \quad (3)$$

where C is the set of all classes, P(cj) is the probability of a class j, w_i is the word occurring in a class j, count(w,c) is the frequency of word w occurring in a class, count(c) is the total number of words that occurred in a class, and V is the set of all unique words of all classes.

The NB algorithm is a supervised machine learning algorithm that encompasses training and testing stages. The training stage aims to use the samples of previously classified data, to generate a training file facilitating the processing of the unclassified documents. To evaluate the suggested system performance, three metrics are used: precision, recall, and F1-measure [22, 23] as expressed in Eqs. (4–6).

$$Precision(P) = \frac{TP}{(TP+FB)} \quad (4)$$

$$Recall(R) = \frac{TP}{(TP+FN)} \quad (5)$$

$$F1-measure = \frac{2(Precision * Recall)}{Precision + Recall} \quad (6)$$

The output of the NB classifier is represented by a confusion matrix that displays the number of documents assigned to each class. Some documents may be assigned correctly, while others are misclassified as illustrated in Table IV. To reduce the variability and make the results generalized, k-fold cross-validation method is used in this research, where K is set to 10 corresponding to the precedent established in prior research [14, 24, 25].

The output of the NB classifier is represented by a confusion matrix that displays the number of documents assigned to each class. Some documents may be assigned correctly, while others are misclassified as illustrated in Table IV.

TABLE IV
CONFUSION MATRIX

		ACTUAL CLASS	
		Class (C)	**Class (Not C)**
Predicted Class	**Class C**	TP	FB
	Class Not C	FN	TN

To reduce the variability and make the results generalized, k-fold cross-validation method is used in this research, where K is set to 10 corresponding to the precedent established in prior research [14, 24, 25]. Results from the BoWs method and the proposed weighting method are displayed in Tables V and VI, respectively.

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE V
RESULTS OF BBC DATASET USING BOWS

10-Fold	F1-Measure
K1	0.7000
K2	0.6287
K3	0.7510
K4	0.7086
K5	0.6932
K6	0.5630
K7	0.7619
K8	0.7139
K9	0.6978
K10	0.5824
Average	**0.6801**

TABLE VI
RESULTS OF BBC DATASET USING PROPOSED WEIGHTING METHOD

10-Fold	F1-Measure
K1	0.7300
K2	0.7157
K3	0.7802
K4	0.7077
K5	0.7701
K6	0.6224
K7	0.7582
K8	0.7053
K9	0.7495
K10	0.6728
Average	**0.7212**

The use of the proposed weight scheme on the BBC Arabic dataset improves the classification results compared with those from the statistical BoWs method. The average F1-measure (**0.7212**) of the scheme outperforms that of the BoWs method, for which the average F1-measure is (**0.6801)**. The justification behind this result is the enhancement of the training file with new features constructed from using relations available in the AWN. However, as mentioned, the inadequacy of AWN with regard to covering all words in the Arabic language indicates that the relations of AWN cannot be entirely relied upon. Therfore, AWN need to be enhanced for further ATC accuracy inprovement.

V. CONCLUSION AND FUTURE WORK

In this work, we propose a new approach for the Arabic TC with the aid of the AWN thesaurus as a lexical and semantic source. With this approach, a weighting scheme is suggested to assign different weights to each relation based on the frequency of the relation in the AWN and the corpus documents corpus. The BoWs of the training documents and their corresponding weighted related terms extracted from the AWN are used as features in the supervised learning of an NB classifier. Results showed that the suggested approach outperformed the statistical BoWs approach. Future work includes selecting words that belong to the same class (i.e., in Table III, we select the words of relations that belong only to WN or MDN). Another idea is to apply a light stemmer on the BBC dataset and then adopt the proposed weighting method.

REFERENCES

[1] Suhad A. Yousif, Venus. W. Samawi, I. Elkabani, and R. Zantout, "Enhancement of Arabic Text Classification Using Semantic Relations with Part of Speech Tagger," W transactions Advances In Electrical And Computer Engineering, 2015, pp. 195-201.

[2] I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig, and N. A. Mahyoub, "Automatic Arabic text categorization: A comprehensive comparative study," Journal of Information Science, vol. 41, 2015, pp. 114-124.

[3] M. S. Khorsheed and A. O. Al-Thubaity, "Comparative evaluation of text classification techniques using a large diverse Arabic dataset," Language resources and evaluation, vol. 47, 2013, pp. 513-538.

[4] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," International Journal of Machine Learning and Cybernetics, vol. 1, 2010, pp. 43-52.

[5] E. Alaa, "A comparative study on arabic text classification," Egypt. Comput. Sci. J, vol. 2, 2008.

[6] R. M. Duwairi, "Arabic text categorization," Int. Arab J. Inf. Technol., vol. 4, 2007, pp. 125-132.

[7] A. Alahmadi, A. Joorabchi, and A. E. Mahdi, "Combining Bag-of-Words and Bag-of-Concepts representations for Arabic text classification," 2014.

[8] A. Karima, E. Zakaria, T. G. Yamina, A. Mohammed, R. Selvam, and V. VENKATAKRISHNAN, "Arabic text categorization: a comparative study of different representation modes," Journal of Theoretical and Applied Information Technology, vol. 38, 2012, pp. 1-5.

[9] Suhad A. Yousif, V. W. Samawi, I. Elkabani, and R. Zantout, "The effect of combining different semantic relations on Arabic text classification," World Comput. Sci. Inform. Technol. J, vol. 5, 2015, pp. 12-118.

[10] F. Harrag and E. El-Qawasmah, "Neural Network for Arabic text classification," in Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the, 2009, pp. 778-783.

[11] M. Sahlgren and R. Cöster, "Using bag-of-concepts to improve the performance of support vector machines in text categorization," in Proceedings of the 20th International Conference on Computational Linguistics, 2004, p. 487.

[12] S. A. Yousif, V. W. Samawi, I. Elkaban, and R. Zantout, "Enhancement of Arabic text classification using semantic relations of Arabic WordNet," Journal of Computer Science, vol. 11, 2015, p. 498.

[13] M. K. Saad and W. Ashour, "Osac: Open source arabic corpora," in 6th ArchEng Int. Symposiums, EEECS, 2010.

[14] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys (CSUR), vol. 34, 2002., pp. 1-47

[15] S. Doan and S. Horiguchi, "An efficient feature selection using multi-criteria in text categorization for naive Bayes classifier," WSEAS Transactions on Information Science & Applications, vol. 2, 2005, pp. 98-103.

[16] M. Alkhalifa and H. Rodríguez, "Automatically extending NE coverage of Arabic WordNet using Wikipedia," in Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco, 2009.

[17] M. M. Boudabous, N. C. Kammoun, N. Khedher, L. H. Belguith, and F. Sadat, "Arabic WordNet semantic relations enrichment through morpho-lexical patterns," in Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on, 2013, pp. 1-6.

[18] T. Brasethvik and J. A. Gulla, "Natural language analysis for semantic document modeling," Data & Knowledge Engineering, vol. 38, 2001, pp. 45-62.

[19] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," Expert Systems with Applications, vol. 36, 2009, pp. 5432-5435.

[20] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB," Int. Arab J. e-Technol., vol. 2, 2011, pp. 124-128.

[21] M. El Kourdi, A. Bensaid, and T.-e. Rachidi, "Automatic Arabic document categorization based on the Naïve Bayes algorithm," in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, 2004, pp. 51-58.

[22] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," Journal of Machine Learning Research, vol. 2, 2002, pp. 419-444.

[23] G. Forman, "An extensive empirical study of feature selection metrics for text classification," Journal of Machine Learning Research, vol. 3, 2003, pp. 1289-1305,.

[24] T. Mullen and N. Collier, "Sentiment Analysis using Support Vector Machines with Diverse Information Sources," in EMNLP, 2004, pp. 412-418.

[25] M. Hadni, S. A. Ouatik, and A. Lachkar, "Effective arabic stemmer based hybrid approach for arabic text categorization," International Journal of Data Mining & Knowledge Management Process, vol. 3, 2013, p. 1-14.

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)